AHRQ Grant Final Progress Report

Title of Project:

Developing Peer-to-Peer Learning Tools for Critical Care Physicians: Peer- and Competency-based Ongoing Approach for Critical Healthcare Evaluations of Skills (P-COACHES)

Principal Investigator

Michael Rosen, MA, PhD – Associate Professor,
The Armstrong Institute for Patient Safety and Quality,
Department of Anesthesiology and Critical Care Medicine,
The Johns Hopkins University School of Medicine,
Joint Appointments in the School of Nursing and the Bloomberg School of Public Health Department of Health Policy and Management

Team Members

Aaron Dietz, PhD – Health System Specialist, U.S. Department of Veterans Affairs

Sadaf Kazi, PhD – Human Factors Specialist, The Armstrong Institute for Patient Safety and Quality, The Johns Hopkins University School of Medicine

Erin Hanahan-Gatz, MPH, Research Program Manager, The Armstrong Institute for Patient Safety and Quality, The Johns Hopkins University School of Medicine

Organization: The Armstrong Institute for Patient Safety and Quality, The Johns Hopkins University School of Medicine

Inclusive Dates of Project: 05/01/2016-10/31/2018

Program Official: Barbara Bartman, barbara.bartman@ahrq.hhs.gov

Grants Management Specialist: Kathryn Carr, kathryn.carr@ahrq.hhs.gov

Acknowledgment of Agency Support

Grant Award Number: 1R03HS024591-01 Revised

Structured Abstract

Purpose: To develop a peer-to-peer assessment system that is theory driven, context appropriate, and psychometrically sound to evaluate teamwork-related skills and patient-centered performance and to develop mechanisms to facilitate adoption and use for peer coaching and feedback.

Scope: This study was conducted on teamwork competencies in intensive care unit (ICU) physicians.

Methods: A systematic review of the literature was conducted to investigate existing methods of conducting peer assessments for physicians. The peer assessment tool was developed through a multistaged approach. First, critical incident interviews were conducted with physicians to identify teamwork competencies that are amenable to observations and measurement during daily rounds. Second, the Delphi method was used to establish consensus among ICU physicians about tool items. Third, a generalizability study was conducted to establish systematic sources of variance in utilizing the tool to conduct peer assessments.

Results: We found 53 studies that had used peer assessment systems to investigate technical and nontechnical competencies in physicians. The most common area in which assessments were conducted was Internal Medicine and Family Medicine. Although a handful of studies developed tools to conduct assessments, the psychometric properties of these tools were not clearly established. We developed a 38-item tool to assess physician competencies in four dimensions: plan of care, team management and norms, teaching and feedback, and patient/family interactions. G study analysis indicated a fair level of measure dependability (G coefficient of .6) and guidance on improving future iterations of peer assessment tools, including appropriate numbers of raters and rating items to include in measurement systems.

Key words: peer assessment, tool development, competency assessment

Purpose

Care of the critically ill requires a multidisciplinary team.[1,2] ICU team coordination has a significant impact on the outcomes and experiences of patients.[3] When a critical care attending physician (i.e., an intensivist) leads daily rounds in the ICU, there is significant benefit to patients and resource utilization.[4] In addition to skills related to diagnosis and treatment planning, intensivist competencies related to adaptive skills, such as teamwork and patient interaction, underlie their ability to provide safe and effective care delivery.[5] Although critical for safety, adaptive competencies are not always formally defined, systematically developed, assessed, or monitored over time.[6-8] Our project addressed this gap by developing and validating a peer-to-peer (P2P) assessment system that can be used to evaluate teamwork and patient-centered performance for intensivists during daily rounds.

Study Aims

There were three specific aims of this project:

> Aim 1: To synthesize the state of science and practice surrounding the use of P2P assessment in healthcare, identify gaps and barriers in current implementation practices, and generate evidence-based guidelines for developing P2P assessment tools;
> Aim 2: To develop and test the feasibility of a prototypical approach for P2P assessment in critical care using novel psychometric analysis techniques; and
> Aim 3: To generate resources to facilitate the implementation of peer evaluation and feedback systems.

Scope

Background

Approximately a third of errors and near misses in critical care can be attributed to teamwork breakdowns.[9,10] The process of daily rounds is central to the care of critically ill patients in an ICU. Rounds provide an opportunity for the entire care team to communicate and reach shared expectations of treatment plans. This team can include physicians, nurses, respiratory and physical therapists, pharmacy, nutrition, social work, patients and their family members, and a variety of other disciplines, depending on the case. Furthermore, daily rounds that are led by an intensivist have been associated with shorter lengths of stay, reduced hospital costs, and a lower likelihood of complications after certain surgical procedures.[4] In the rounding process, the ICU team examines patients, reviews pertinent data, and collaborates to formulate a care plan. The intensivist leads this team on rounds and is ultimately responsible for guiding the rounds process and creating the daily plan of care. Furthermore, rounds often have additional goals that may include education, team building, event analysis (e.g., review of cases and possible errors), and addressing the needs of patients and their loved ones.

Context

The intensivist plays a pivotal role in care delivery (e.g., management of team resources), medical education (e.g., foster development of clinical skills), and the establishment of behavioral norms (e.g., role modeling). Learners (e.g., residents) will likely adopt what they perceive to be desirable and acceptable behavior in actual practice. Unfortunately, current P2P tools are underdeveloped and have yet to demonstrate the reliability and validity evidence necessary for implementation. Finally, indicators of teamwork and patient-centered performance are conspicuously absent from current approaches, and no method exists specifically to evaluate proficiency in critical care.[7,11] Therefore, the proposed research seeks to develop a P2P assessment system that is theoretically motivated and psychometrically validated to evaluate teamwork-related skills and patient-centered performance and to develop mechanisms to facilitate adoption and use for peer coaching and feedback.

Study design

Project aims were addressed in three related studies: 1) a systematic literature review, 2) a Delphi study, and 3) a Generalizability (G) study. Design and methods for each study are described below.

Systematic Literature Review of Peer Assessment Systems Methods

The primary questions guiding our systematic review of the literature on physician peer assessments were:

1. What is the purpose for which peer assessments were conducted?
2. What clinical domains are peer assessments conducted in?
3. Do researchers establish psychometric properties of peer assessment tools?

We conducted a systematic review of peer-reviewed journal articles on PubMed using search terms related to peers, assessments, and physician, which resulted in 1162 hits. After title review, we retained 472 articles, which was further reduced to 86 articles after abstract review. We screened these 86 articles for inclusion through full-text coding.

Coding. Two coders independently coded 10 of the 86 articles until 80% coding consensus was established. The remaining 76 articles were then split between the two coders and single coded. Articles were coded on the role of the peer assessor (e.g., attending physician, fellow physician, etc.) as well as on the individual being assessed, the purpose underlying the review (e.g., skill assessment, compliance assessment), the clinical domains in which peer reviews were conducted, and description of the process of conducting peer reviews. After full-text coding, we retained a final set of 53 articles.

Delphi Study for Peer Assessment Tool Development

Participants. Two rounds of Delphi[12] surveys were conducted over a period of 3 months. Both surveys were conducted on the same sample of nine participants (seven critical care attending physicians; two critical care fellows) who worked in the intensive care units at Johns Hopkins hospital. Participants received monetary compensation for completing each round of the Delphi survey.

Materials and procedure. We developed a tool to assess important nontechnical competencies during multidisciplinary patient-centered rounds in the ICU based on a technical review of the literature about P2P assessments in healthcare and critical interviews with confederate ICU physicians and nurses.[13] The P2P tool was organized around four behavioral dimensions representing nontechnical competencies during multidisciplinary rounds: plan of care, team management and norms, teaching and feedback, and patient/family interactions (see Table 1 for definition of each dimension). Each dimension was associated with numerous positive and negative behavior markers. Behavioral markers represent indicators of effective and ineffective performance that can be overtly observed and assessed.[14,15]

Table 1. Definitions of the Four Dimensions of the P2P Rounding Tool

Dimension Name	**Definition**
Plan of care	Reviewing and analyzing relevant case information and treatment planning
Team management and norms	Managing team resources and personnel while ensuring an inclusive atmosphere and modeling desired behaviors
Teaching and feedback	Providing opportunities to foster the development of knowledge and skills for trainees involved in the presentation and treatment planning for the current case
Patient/family interactions	Incorporating the patient and/or their loved ones during rounds

The Delphi survey was developed and administered online. In Round 1, participants rated 44 behavioral markers. Each marker was rated on its importance to patient care and the degree to which it was observable on a five-point rating scale ranging from strongly disagree to strongly agree. Respondents also suggested clarifications in phrasing of behavioral markers, assessed the suitability of the four dimensions, and suggested other dimensions to include in the tool. In Round 2 of the Delphi survey, participants evaluated modifications to a subset of markers based on feedback from Round 1.

Generalizability (G) Study Design

Generalizability (G) theory was applied to provide evidence of the assessment system's reliability and construct validity.[16] G theory is a novel psychometric approach that applies an analysis of variance (ANOVA) to quantify systematic variance from multiple sources simultaneously.[17-19] G theory was applied in the present study to evaluate the potential utility of the P2P assessment system during daily rounds to (1) differentiate performance and (2) establish that the competencies identified in the formal needs analysis are unique. The tool developed through the Delphi process described above was used to collect data for evaluation. Specifically, two raters scored 2 days of weekday morning patient rounds for each of four ICU attending physicians.

Table 2. Sources of Variance for G Study

Source of Variation	**Description**

Intensivist (I)	• Systematic variability in ICU intensivist performance
Rater (R)	• Variance attributable to raters across intensivists, competencies, and observations.
Competencies (C)	• Systematic variability in how competencies are rated across intensivists and observations
Observations (O:I)	• Systematic variability in a specific intensivist's performance across measurements
I X R	• Systematic variability in how raters score a particular intensivist
(O:I) X R	• Systematic variability in how raters score specific observations
I X C	• Variance due to intensivists performing differently on certain competencies than others
(O:I) X C	• Systematic variability in how competencies are rated by observation
O:I X R X C, e	• Residual error (unexplained variance) [Note: this term is indistinguishable from RCI.]

Results

Literature Review Principal Findings

1. Studies of peer assessments were reported in 53 journal articles.
2. Assessments were overwhelmingly conducted by attending physicians (n = 37, 70%) and assessed the performance of attending physicians (n = 33, 62%).
3. A majority of assessments (n = 31, 58%) used medical records as the source of information about physician competencies.
4. Most articles (n = 30, 57%) described processes for conducting peer reviews. Only eight articles (15%) focused on development of novel tools for assessing peers. The remaining articles used previously validated tools to assess physician peers.
5. The most common clinical domain in which peer reviews were conducted was internal medicine (n = 14, 26%), followed closely by family medicine (22%) and then by surgery (n = 8, 15%) and radiology (n = 5, 9%). Four articles (7%) each conducted assessments in emergency medicine and pediatrics. Three articles (6%) each conducted peer assessments in radiation oncology, general medicine, obstetrics/gynecology, psychiatry, and primary care. Two articles (4%) conducted peer assessments in anesthesia. Finally, the domains of oncology, cardiology, palliative care, geriatrics, neurophysiology, orthopedics, and addiction each accounted for one article.
6. Peer assessments were conducted on a variety of tasks, including review of treatment plans, test ordering, record keeping, etc.
7. There were several commonalities in the process of conducting peer reviews, including generating items or criteria for the review and obtaining data for conducting the review, either through a review of medical records or behavioral observations, followed by immediate recording of peer review results. However, there were wide variations in the process that influenced the psychometric properties of the tool, including review of tool criteria by expert samples, providing training to raters in using the peer review process, providing feedback to individuals who were reviewed, and establishing reliability of peer assessment (e.g., through generalizability studies).

8. Although there is variety in the tools that exist to evaluate physician competencies, there is a need to establish psychometric properties of these assessment instruments.

Delphi Study Principal Findings

Round 1:

1. A marker was labeled important/observable if at least 75% of respondents marked it above the midpoint. Table 3 shows the classification of markers from Round 1 of Delphi. Eighty-four percent (n = 37) of the markers were judged important and observable. Of these, the phrasing of the remaining 25% (n = 11) was judged as sufficiently capturing its relevant marker.
2. Participants had suggestions to improve the wording of 59% (n = 26) of markers that were judged important and observable. Based on a thematic analysis of the comments, we created modified phrasing for these markers for Round 2 of the Delphi survey.
3. Sixteen percent (n = 7) of markers were judged as possessing low observability. Of these, only one marker was also judged important. Five markers, including the one judged important, had suggestions for rephrasing and were modified for Round 2.

Table 3. Classification of Markers Above the Midpoint from Round 1 of Delphi

		Dimension Name				Total
		Plan of care	Team management and norms	Teaching and feedback	Patient/family interactions	
Survey response classification	High importance & high observability; no edits suggested	0	3	4	4	11
	High importance & high observability; with edits	8	10	1	8	26
	High importance & low observability; with edits	1	0	0	0	1
	Low importance	2	2	2	0	6
Total		11	15	7	11	44

Round 2: In this round, the Delphi survey primarily focused on markers that were judged important in Round 1 but had suggestions to modify the phrasing. Specific response options for such markers depended on their classification in Round 1.

1. For the 26 markers judged important and observable, participants chose between either retaining the original phrasing or retaining the marker with the new phrasing. Similar to Round 1, we calculated the percent of responses obtained for each option. If no response option was selected by at least 75% of the respondents, then we deleted the item. This resulted in deletion of four items.

2. For the single marker judged important but unobservable from Round 1, but which had suggestions for clarifying phrasing, and for the four markers judged unimportant from Round 1, participants could choose between three response options: (1) retaining the original phrasing; (2) retaining the new phrasing; (3) deleting the item. Three items did not reach a 75% consensus and were subsequently deleted.
3. Table 4 shows the final set of markers after Round 2 of the Delphi survey. The dimensions of plan of care and team management and norms are relatively balanced between positive and negative markers; this is less true for patient/family interactions. Future research is needed on markers to capture appropriate teaching and feedback during rounds.

Table 4. Positive and Negative Markers Associated with the Four Dimensions of the P2P Tool after Round 2 of Delphi

Plan of care: Review and analysis of relevant case information, and treatment planning

Positive	Negative
• Team members assess the patient's condition, effectiveness of current management strategies, and plan of care. • Team members discuss anticipated patient progress, including progress toward discharge home or transfer, and discuss diagnosis and prognosis. • Team explicitly discusses daily goals and targets while discussing individual systems or after all systems have been discussed. • Team identifies action items (e.g., procedures) and resources needed, both inside and outside the unit, (e.g., consultants, equipment, etc.) or any rate-limiting steps to achieving action items.	• Communication of patient assessment and management plans is inadequate, lacking important details. • Team does not discuss anticipated barriers to daily goals, alternative management strategies, or specific criteria for decisions to be made based on patient's response to treatments • Daily goals of care are inadequately communicated to the entire team. • Team does not identify action items (e.g., specific procedures) and resources needed (both inside and outside the unit; e.g., consultants, equipment, etc.) or any rate-limiting steps to achieving action items. • There is no confirmation that orders discussed during rounds have actually been placed and documented or no clarification on who will place orders.

Team Management and Norms: Managing team resources and personnel while ensuring an inclusive atmosphere and modeling desired behaviors

Positive Markers	Negative Markers
• Attending physician/fellow acknowledges good work and provides positive reinforcement. • Attending physician/fellow facilitates the flow of conversation to ensure the timely completion of all relevant discussion concerning a patient. • Attending physician/fellow seeks input from all team members and encourages questions. • Attending physician/fellow assigns team member to review rounding discussion with team members who may be unable to participate (e.g., nurse, respiratory therapist, pharmacist, other consult, etc.). • Multidisciplinary team members (e.g., nursing, pharmacy, respiratory therapy) actively participate in providing updates and input for treatment planning.	• The attending physician/fellow does not take all reasonable actions to ensure a nurse representative is present before the case is discussed. • Nurse does not participate in rounds. • Input from team members is dismissed without explaining the reason for the dismissal. • The input of multidisciplinary team members (e.g., nursing, pharmacy, respiratory therapy, consultants) is not sought during discussion of relevant case information. • Multiple team members are distracted (e.g., checking emails, texts) or are talking simultaneously. • Conversation unrelated to the current case results in disorganized rounds. • Team discusses tasks but does not discuss or resolve the assignment of task responsibilities.

Teaching and Feedback: Providing opportunities to foster the development of knowledge and skills for trainees involved in the presentation and treatment planning for the current case

Positive Markers	Negative Markers
• Elicits evidence for treatment and missing variables • Balances teaching with timeliness of rounds • Provides feedback on important points and areas of omission after case presentations • Allows time/opportunities for discussion and for team members to ask questions	• Does not provide opportunities for questions or discussion • Uses intimidation/fault finding as feedback mechanism • Feedback about patient care decisions or suggested plan of care or presentation style is not provided or is insufficient.

Patient/Family Interactions: Incorporating the patient and/or their loved ones during rounds

Positive Markers	Negative Markers
• Team members offer introductions and invite patients and/or their family/ visitors to participate in rounds. • Ensures that patients and their family members/visitors are aware about the format of rounds before beginning rounds • Empowers the patient or their family members/visitors to speak up and ask questions • Attending physician/fellow uses lay terms to ensure that the patient and/or their family/visitors understand the discussion or assures the patient and/or their family/visitors that they will return after rounds to answer questions. • Tries to arrange for an interpreter for non-English speaking patients and their family members/visitors during rounds or ensures that obtaining a translator is incorporated in the plan of the day • Appropriately manages conflict and negotiates criticism from patients or their family members/visitors	• Does not engage patients or their family members/visitors during rounds • Does not empower the patient and/or their family members/visitors to speak up and ask questions • Does not adequately respond to or address questions from the patient or their family members/visitors during rounds • Interrupts or deflects questions from the patient or their family members/ visitors during rounds without offering to return and discuss questions in more detail after rounds

Scoring guide

Score	Interpretation
1	Poor: Performance was expected but not observed; performance consistently demonstrated negative behaviors.
2	Marginal
3	Neutral/Acceptable: Performance was adequate. Attending demonstrated positive behaviors but also showed areas for improvement. Attending competency was acknowledged, but opportunities to further demonstrate competency were precluded due to patient conditions or situation.
4	Good
5	Very Effective: Performance consistently demonstrated positive behaviors throughout the entire observation.
NA	Performance was not expected for this particular round.

G Study Principal Findings

A G study was performed on data generated by two raters observing a total of 48 patient rounding events sampled from four attending physicians. The rating tool refined in the Delphi study above included four subdimensions. However, the patient and family subdimension was difficult to observe in practice, as family members were not often present during rounds. G study analyses are based on ANOVA procedures and require balanced designed (i.e., no missing data). Therefore, only three subdimensions were included in the analysis, resulting in a total of 336 ratings. The observation and estimation designs for the G study are detailed in Table 5.

Table 5. G Study Observation and Estimation Designs

Facet	Label	Observed Levels	Universe Levels
Intensivist	I	4	INF (random)
Rater	R	2	INF (random)
Competencies	C	3	3 (fixed)
Observations	O:I	14	INF (random)

In addition to the observed levels specifying the levels of a facet present in the actual data, G study analyses require specification of desired levels for the universe score. This represents the degree to which measurements are intended to be generalized beyond the observed data. Specifying a universe level as fixed (i.e., the same as the observed levels) means that scores are not intended to be generalized past the observed levels for that facet. Specifying a random universe level means that it is desired to generalize the ratings to as broad a population of levels as possible. In this design, intensivists, raters, and observations are specified as random, and competencies are specified as fixed.

Table 6 provides the detailed ANOVA results, and Table 7 details the G study results, in which variance components are grouped into differentiation and instrumentation facets.

Table 6. ANOVA Results for Facets

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
I	6.42	3	2.14	-0.01	0.00	0.00	0.1	0.02
R	1.19	1	1.19	-0.01	0.00	0.00	0.0	0.01
C	13.20	2	6.60	0.03	0.03	0.02	3.1	0.04
O:I	65.48	52	1.26	0.10	0.14	0.14	19.7	0.04
IR	3.71	3	1.24	0.01	0.02	0.02	2.7	0.02
IC	9.49	6	1.58	0.02	0.02	0.02	3.0	0.03
RC	3.79	2	1.90	0.02	0.02	0.02	2.9	0.02
RO:I	20.43	52	0.39	0.06	0.13	0.13	17.8	0.03
CO:I	48.31	104	0.46	0.13	0.13	0.13	17.5	0.03
IRC	4.23	6	0.71	0.04	0.04	0.04	4.8	0.03
RCO:I	21.64	104	0.21	0.21	0.21	0.21	28.4	0.03
Total	197.89	335					100%	

Note: I = Intensivist; R = Rater; C = Competencies; O:I = Observations nested in Intensivists

Table 7. G Study Table (Measurement Design ICO/R)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
I	0.00		.....			
	.....	R	.....		(0.00)	0.0
C	0.02		.....		.....	
O:I	0.14		.....		.....	
	.....	IR	0.01	4.8	0.01	4.8
IC	0.02		.....		.....	
	.....	RC	0.01	5.1	0.01	5.1
	.....	RO:I	0.07	31.5	0.07	31.5
CO:I	0.13		.....		.....	
	.....	IRC	0.02	8.5	0.02	8.5
	.....	RCO:I	0.10	50.0	0.10	50.0
Sum of variances	0.32		0.21	100%	0.21	100%
Standard deviation	0.56		Relative SE: 0.46		Absolute SE: 0.46	
Coef_G relative	0.60					
Coef_G absolute	0.60					

Note: I = Intensivist; R = Rater; C = Competencies; O:I = Observations nested in Intensivists

As illustrated in Figure 1, the facets of differentiation (i.e., sources of desirable variance) accounted for large portions of the overall variance in scores. The competency by instances of teamwork (CO:I) accounted for 17.5% of the variance, and the main effects for instances (O:I) and competencies (C) accounted for 19.7% and 3.1% of variance in ratings, respectively. Small proportions of variance were accounted for by the main effect of the intensivist (I; 0.1%) and the intensivist-by-competency (IC; 3%) interaction. Taken together, these facts of differentiation account for large proportions of overall rating variance. However, unaccounted-for variance (RCO:I at 28% and IRC at 5%) and variance due to facets of instrumentation (i.e., undesirable sources of variance) were present as well. There was no variance due to a main effect for rater (R: 0%); however, there were was large variance associated with an interaction between raters and observations of teamwork (RO:I) and small variance associated with an intensivist-by-rater (IR; 3%) interaction. The overall G coefficient (which can be interpreted as a reliability coefficient, using a heuristic cutoff of .8 and above as substantial dependability) was .6. This indicates that the dependability of the measures should be improved. To further investigate how this tool and measurement process could be improved, we conducted a Decision (D) study.

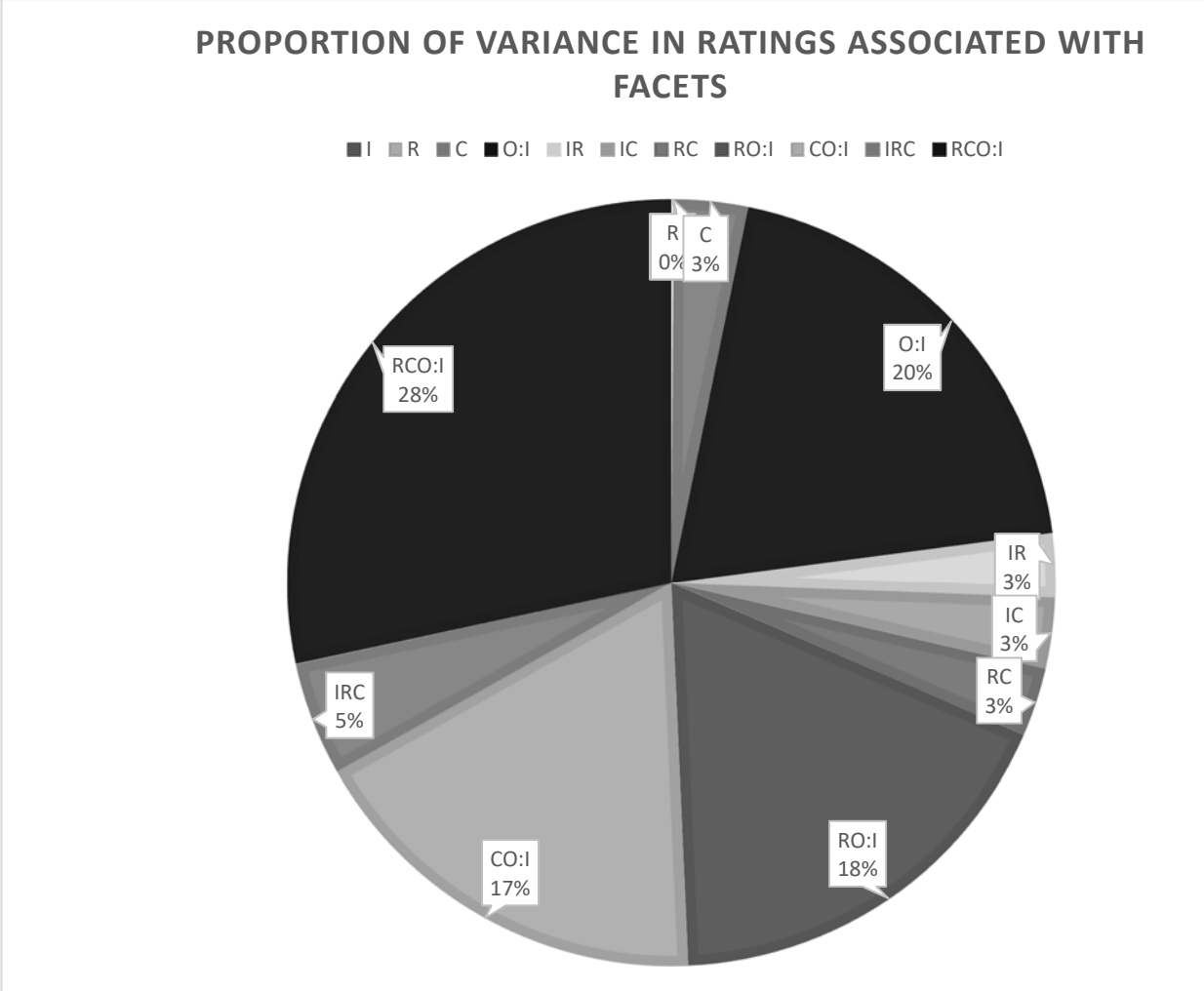

Figure 1. Illustration of proportion of variance in ratings attributable to facets of measurement

A Decision (D) study analysis was conducted using the above detailed G study information. The purpose of a D study is to extrapolate from existing data to guide future development of measurement systems. Specifically, D studies involve an optimization process in which different model parameters are varied to explore the impact on G coefficients. We explored changes to rater and competency facets, as illustrated in Figure 2.

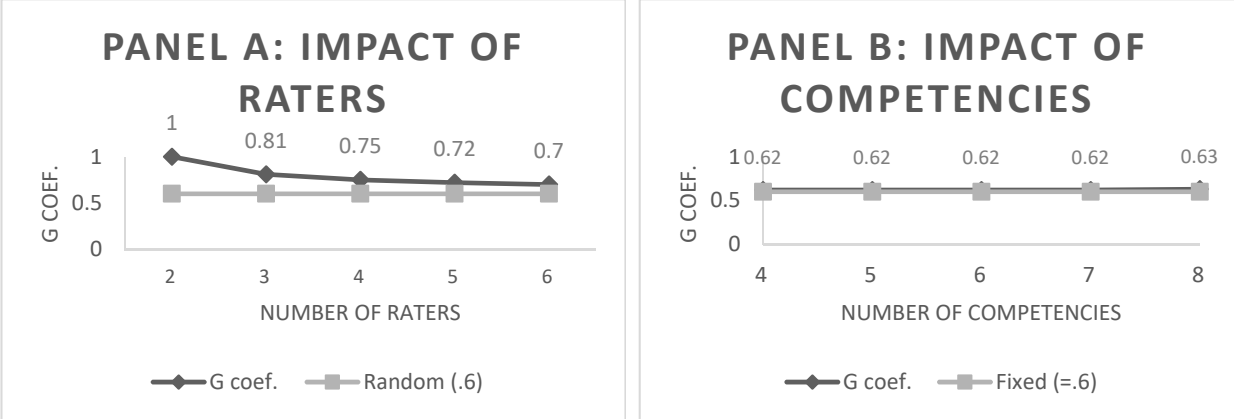

Figure 2. *Panel A* illustrates the effects of varying the number of raters we wish to generalize to in future data collections. The G study model specified the universe parameter for raters as random, meaning we wished to generalize to the largest possible set of raters. The G coefficient was .6 at this level. We varied the universe levels for 2 (a fixed facet, indicating no intention to generalize past the observed data) and 6. *Panel B* illustrates the effects of varying the number of competencies we wish to generalize to. The G study included competencies as a fixed facet (a universe level of 3, indicating no intention to generalize past observed data). We varied universe levels for competencies from 4 to 8.

The D study findings illustrated above indicate that 1) constraining the set of raters dramatically improves the predicted G coefficient, and 2) adding additional competencies modestly increases the G coefficient. Based on these findings, an 'optimal' model was specified using three levels for the rater universe score and four levels for the competencies universe score. This resulted in a G coefficient of .82.

Discussion

The literature review provided insights into the state of the science and practice around peer assessment systems for attending physicians:

- **There are few details in the literature on assessment tools or processes**. These systems exist, but evaluation of and reporting on their structure and process are needed.
- **Most peer assessments focus on technical, not teamwork, competencies**. Athough technical competencies are no doubt important, there are fewer opportunities to formally develop teamwork-related competencies.

The Delphi study generated behavioral markers related to nontechnical competencies during patient rounds:

- **Attending physicians gained consensus on markers for four dimensions of teamwork and patient- and family-centered care** that are essential during daily patient rounds: plan of care, team management and norms, teaching and feedback, and patient/family interactions.
- Although we were successful in obtaining a variety of behavioral markers associated with the dimensions of plan of care, team management and norms, and patient/family interaction, **future work should explore expansion of markers for teaching and feedback**.

G and D study findings provide an assessment of the tool's properties and a path forward for future development and refinement of peer assessment practices:

- **Facets of differentiation generated large portions of rating variance**. Specifically, high levels of variance associated with observations, competencies, and an interaction between the two indicate a good structure to the tool. Competencies were rated differently for each instance of patient rounding, demonstrating that the subdimensions represent unique aspects of teamwork.
- **There were low levels of variance associated with intensivists**. Taken with the findings above, this indicates that there was more variance within attending physicians than between. This could be due to the relatively small number of physicians or the fact that they were all from the same department. This creates challenges to the use of the tool for individual feedback to a physician. However, the large proportion of variance associated with the interaction between observations and competencies indicates value in using the tool for developmental feedback.
- **Systematic effects of raters by observations of teamwork indicate a need for increased rater training**. Although there was no overall systematic variance associated with raters, there were differences in raters by observations. This means that, for certain patient rounding events, one rater would have systematically higher scores than another. This bias was present for only certain events and not overall.
- **Patient and family engagement behaviors are difficult to observe during rounding events**. Although the behavioral markers for this subdimension were rated observable by attending physicians in the Delphi study, they seldom occurred in actual rounding. The makers all described interactions with patient or family members present during rounds. This happened infrequently; however, teams did discuss patient and family member engagement strategies without their presence more frequently (e.g., discussing which family members have been involved or are decision makers, plans to hold family meetings or approaches to communicating with different family members). We plan to modify the marker system to include these types of interactions.
- **Future peer assessment systems should focus on small groups of assessors and larger sets of competencies.** The D study indicated that measurement system dependability would increase if the pool of potential raters was constrained and if more competencies were included in the assessment. We believe that modification of the above patient and family engagement competency will address this shortcoming of the tool.

Conclusions

The peer-to-peer rating tool developed in this work was demonstrated to have consensus agreement from attending physicians about the importance of teamwork behaviors in rounding. Perceived and actual observability for patient and family engagement behaviors differed substantially. G study results indicate that the tool has fair dependability, and D study analyses highlight the importance of maintaining a small group of peer raters and ensuring that at least four competency ratings are used.

Significance

The literature review highlights the paucity of tools available to structure peer-to-peer feedback on teamwork behaviors. This study contributes a practical tool and robust psychometric evaluation of generalizability and dependability of ratings generated using the tool.

Implications

Peer-to-peer assessment systems for critical care attendings can include dependable measures of teamwork competencies. Capturing dependable ratings is challenging, and raters are important sources of systematic variance and random error. To generate dependable measurements, peer rating tools should focus on developing a small, core group of raters and should use tools incorporating at least four subdimension ratings.

List of Publications

Presentation:

Kazi, S., Dietz, A., Berenholtz, S., Sapirstein, A., & Rosen, M.A. (2018). *Developing a Peer-to-Peer Learning Tool for Critical Care Physicians.* Paper presented at the 62nd Annual Meeting of the Human Factors and Ergonomics Society, Philadelphia, PA.

Publications:

We are currently preparing two manuscripts to be submitted to journals targeting critical care medicine and physician competency development.

1. Kazi, S., Dietz, A., Berenholtz, S., Sapirstein, A., & Rosen, M.A. Developing a validated instrument to assess nontechnical competencies in physicians. *Critical Care Medicine.*
2. Kazi, S., Hanahan-Gatz, E., Dietz, A., & Rosen, M.A. A systematic review of physician peer assessment systems. *Academic Medicine.*

REFERENCES

1. Dietz AS, Pronovost PJ, Mendez-Tellez PA, et al. A systematic review of teamwork in the intensive care unit: What do we know about teamwork, team tasks, and improvement strategies? J Crit Care. 2014;29(6):908-914.
2. Reader TW, Flin R, Mearns K, Cuthbertson BH. Developing a team performance framework for the intensive care unit. Crit Care Med. 2009;37(5):1787-1793.
3. Weaver SJ, Rosen MA, Salas E, Baum KD, King HB. Integrating the science of team training: Guidelines for continuing education. J Contin Educ Health Prof. 2010;30(4):208-220.
4. Dimick JB, Pronovost PJ, Heitmiller RF, Lipsett PA. Intensive care unit physician staffing is associated with decreased length of stay, hospital cost, and complications after esophageal resection. Crit Care Med. 2001;29(4):753-758.
5. Davis D, O'Brien MA, Freemantle N, Wolf FM, Mazmanian P, Taylor-Vaisey A. Impact of formal continuing medical education: Do conferences, workshops, rounds, and other traditional continuing education activities change physician behavior or health care outcomes? JAMA. 1999;282(9):867-874.
6. Berenholtz S, Pronovost PJ. Barriers to translating evidence into practice. Curr Opin Crit Care. 2003;9(4):321-325.
7. Evans R, Elwyn G, Edwards A. Review of instruments for peer assessment of physicians. BMJ. 2004;328(7450):1240.
8. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
9. Donchin Y, Gopher D, Olin M, et al. A look into the nature and causes of human errors in the intensive care unit. Crit Care Med. 1995;23(2):294-300.
10. Pronovost PJ, Thompson DA, Holzmueller CG, et al. Toward learning from patient safety reporting systems. J Crit Care. 2006;21(4):305-315.
11. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387-396.
12. Hsu CC, Sandford BA. The delphi technique; making sense of consensus. Practical Assessment, Research, and Evaluation. 2007;12(10):1-8.
13. Flanagan JC. The critical incident technique. Psychological Bulletin. 1954;51(4):327-358.
14. Dietz AS, Pronovost PJ, Benson KN, et al. A systematic review of behavioural marker systems in healthcare: What do we know about their attributes, validity and application? BMJ Qual Saf. 2014b;23(12):1031-1039.
15. Flin R, Martin L. Behavioral markers for crew resource management: A review of current practice. The International Journal of Aviation Psychology. 2001;11(1):95-118.
16. Brennan RL. Generalizability theory. New York, NY: Springer; 2001.
17. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: A key to unlock professional assessment. Med Educ. 2002;36(10):972-978.
18. Crossley J, Russell J, Jolly B, et al. 'I'm pickin' up good regressions': The governance of generalisability analyses. Med Educ. 2007;41(10):926-934.
19. Kraiger K, Teachout MS. Generalizability theory as construct-related evidence of the validity of job performance ratings. Human Performance. 1990;3(1):19-35.